Automated topic prediction of LitCovid using BioBERT

Saipradeep VG, Naveen Sivadasan, Aditya R Rao, Thomas Joseph
TCS Research, Life Sciences, Hyderabad, INDIA

Background: The BioCreative VII- Track 5 challenge aims to increase the accuracy of automated topic prediction in the massively growing COVID-19 literature to help research community find effective diagnostics, drugs and vaccines for COVID-19. The challenge requires the participants to assign each of the 60k articles from LitCovid [1,2] database with upto seven topics such as Diagnosis, Treatment.

Methods : We propose two different approaches for the multi label assignment task. In the following, we outline the different steps involved in these two models.

a) Training : In the first approach, each article in the dataset is split into two parts, namely, the abstract part and the 'rest' part. The abstract part contains only text from the abstract field while the rest consists of the article title, the keywords and the journal type metadata fields. The second approach also splits the article content into two parts. Additionally, it performs NER on the abstract and title texts. Specifically, in the second approach, we repurposed our text-mining framework PRIORI-T [3] to perform NER covering 27 different entity types such as genes, proteins, diseases, chemicals/drugs and non-biomedical entities such as country, non-phrama interventions etc. These annotations are then masked by replacing the tagged entity with its entity type token.

b) *Prediction :* We used the large BioBERT [4] base model pretrained on MNLI for this challenge. These base model was fine-tuned during training. The first approach uses the untagged train and validation datasets, while the second uses the tagged rain and validation datasets for fine-tuning BioBERT. For the first approach, the training resulted in two fine tuned BioBERT models, namely the model trained on the abstracts part and the model trained on the 'rest' part. The final prediction is obtained by performing an ensemble of the prediction outcomes of these two models. Similarly, the second approach also yielded two trained BioBERT models and the final prediction is performed in a similar fashion. We refer to the final models obtained as part of our first approach as *Model 1* and the model obtained as part of the second approach as *Model 2*.

We experimented with three different ensemble approaches, namely, simple average, weighted average and maximum. The instance-based and label-based validation f1-scores showed slightly better performance for maximum ensemble approach. Hence we used this as our final score aggregation approach.

Results: On the test data, Model 1 achieved instance-based f1 score, label-based macro f1 score and label-based micro f1 scores of 0.8845, 0.8495 and 0.7896 respectively. Model 2 achieved scores of 0.8267, 0.8157 and 0.7181 respectively. The baseline model (MLNet[5]) provided by the challenge organizers achieved scores of 0.8678, 0.7655 and 0.8437 respectively.

Discussion and Conclusions: Our Model 1 showed better performance than Model 2 and the challenge baseline model (MLNet). Further, when benchmarked against the 80 team submissions for this challenge, Model 1's label-based macro f1-score was close to the median f1 score and instance-based f1-score was close to mean f1 score. We separately tried data augmentation by including additional articles from MEDLINE for imbalanced label classes such as Epidemic forecasting and Transmission. However, these augmentations did not improve the performance.

References:

1. Chen Q., Allot A., & Lu Z. LitCovid: an open database of COVID-19 literature. *Nucleic Acids Research.* 2021 Jan 8;49(D1):D1534-40.

2. Chen, Q., Allot, A. and Lu, Z., 2020. Keep up with the latest coronavirus research. *Nature*, 579(7798), pp.193-194.

3. Rao A, Joseph T, Saipradeep VG, Kotte S, Sivadasan N, Srinivasan R., PRIORI-T: A tool for rare disease gene prioritization using MEDLINE. *PLoS One*. 2020;15(4):e0231728.

4. Lee, Jinhyuk and Yoon, Wonjin and Kim, Sungdong and Kim, Donghyeon and Kim, Sunkyu and So, Chan Ho and Kang, Jaewoo, BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics,* Volume 36, Issue 4, 15 February 2020, Pages 1234–1240.

5. Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C. and Lu, Z., 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, 26(11), pp.1279-1285.